STAT 306 Group P1 Interim Proposal

Background Description: The Chinese automobile company Geely Auto aims to enter the U.S. market by establishing a local manufacturing facility and producing cars domestically. A high accuracy predictive model can help Geely Auto to determine their pricing strategy based on the specific car features in the competitive U.S. market. This analysis will not only identify which features American consumers value the most but also demonstrate how these features impact car prices. By understanding the influence of car features, Geely Auto is able to price their new cars to match with market expectations, ensuring their entry into the U.S. market is aligned with consumer preferences.

Research Question: What are the key features that contribute to the highest accuracy of car price predictions in a multiple linear regression model using forward selection and cross validation?

Dataset features: The dataset includes 205 rows and 26 columns. It has no missing values and no duplicate values, ensuring data quality for machine learning purposes.
The dataset originally contains 25 covariates, but some of them are redundant or challenging to interpret. For example, features like "car width", "car height", "car length", and "car weight" all represent similar car's attributes. To reduce redundancy, we decided to keep one of these attributes that best represents the car's size. Additionally, features such as "car_ID" and "symboling" are difficult to interpret and do not appear to provide meaningful information about the car's characteristics. These features were deemed irrelevant to our analysis, so we drop these features. After filtering, the dataset now contains 14 relevant features(includes output), improving both interpretability and focus for analysis.

Variable Measured: The dataset was published in 2019. The exact timeframe of data collection is not specified. All features are measured in U.S. customary units, and all data points are from vehicles in the U.S based on various market surveys.

Variable	Type	Description	Unit
fueltype	Binary Categorical	The type of fuel the car uses: "gas" or "diesel"	Categorical
aspiration	Categorical	the type of aspiration used in the car's engine	Categorical
doornumber	Categorical	the number of doors on the car	Categorical
carbody	Categorical	the body style of the car	Categorical
drivewheel	Cateogrical	the drive wheel type of car	Categorical
enginelocation	Categorical	the location of the engine in the car	Categorical
carlength	Numerical	the overall length of the car	inch
enginesize	Numerical	the size of the engine	cubic inches
cylindernumber	Categorical	the number of cylinders in the car's engine	Categorical
fuelsystem	Categorical	the type of fuel system used in car	Categorical
horsepower	Numerical	the power of the engine	hp
peakrpm	Numerical	the engine's maximum revolution per minute at the peak power	peak revolutions per minute
citympg	Numerical	the fuel efficiency in miles per gallon driving in city	miles per gallon
price(output)	Numerical	the price of the car	US dollar

Yu Chang 47945050

Works on feature selection and engineering, identifying which variables should be used in the model and creates any additional features that may improve model performance, also finalizing the report.

Zhuoran Wang 23359425

Prepares the data for analysis, including cleaning, handling missing values, transforming variables, and performing exploratory data analysis (EDA). Perform forward selection based on accuracy metric.

Nishok Rao 44572535

Lead the exploratory analysis component of the project, helping to create visualizations(plots) to summarize characteristics of the data. Responsible for polishing the final report.

Ruksana Tet Toe 17384280

Ensures the data is ready for modeling and is of high quality. Polish the final report. Conclude on the deficiencies and limitations of our model.

Reference:

Kumar, M. (2019). Car Price Prediction Multiple Linear Regression. Kaggle. Retrieved November 3, 2024, from https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/data